REPORT RESUMES

ED 003 349 24

THE USE OF PROGRAMMED INSTRUCTION IN THE DEVELOPMENT OF RECOGNITION SKILLS.

BY- WILLIAMS, JOANNA P.

PENNSYLVANIA UNIV., PHILADELPHIA

REPORT NUMBER CRP-2100 PUB DATE 65

EDRS PRICE MF-$0.18 HC-$3.72 93P.

DESCRIPTORS- *RECOGNITION, PROGRAMING, LINEAR PROGRAMING, *PROGRAMED INSTRUCTION, *READING READINESS, *LEARNING THEORIES, *SKILL DEVELOPMENT, GRADE 6, PHILADELPHIA, PENNSYLVANIA

OBJECTIVES OF THIS RESEARCH WERE TO (1) INVESTIGATE THE EFFECTS OF THE RESPONSE MODE IN PROGRAMED INSTRUCTION, (2) EXTEND THE THEORETICAL PRINCIPLES ON WHICH PROGRAMED INSTRUCTION IS BASED, AND (3) DEVELOP EFFECTIVE METHODS OF TRAINING RECOGNITION SKILL. A SERIES OF FOUR EXPERIMENTS, INVOLVING SIXTH-GRADE PUPILS, WERE REPORTED. IN EXPERIMENT 1, STUDENTS WORKED THROUGH A LINEAR PROGRAM OF 2 VERSIONS, CONSTRUCTED-RESPONSE AND STRAIGHT TEXT. IMMEDIATE AND DELAYED TEST SCORES ON AN OBJECTIVE CRITERION TEST WERE HIGHER FOR THOSE STUDENTS WHO RECEIVED THE CONSTRUCTED-RESPONSE VERSION. EXPERIMENT 2 SHOWED THAT PERFORMANCE ON A CONSTRUCTED-RESPONSE TEST WAS SUPERIOR FOR THE CONSTRUCTED-RESPONSE TRAINING MODE. PREDICTED INTERACTION BETWEEN THE TRAINING MODE AND TYPE OF TEST ITEM WAS SIGNIFICANT. FINDINGS FROM A THIRD EXPERIMENT SUGGESTED THAT USING A COMBINATION OF MODES IN TRAINING MAY LEAD TO CONFUSION AND REDUCE THE EFFECTIVENESS OF A PROGRAM. EXPERIMENT 4 CONCERNED RECOGNITION OF LETTER-LIKE FORMS, ONE OF THE PRIMARY STEPS IN LEARNING TO READ. DATA SUGGESTED THAT A RELATIVELY LARGE AMOUNT OF TIME MIGHT WELL BE DEVOTED TO PRACTICE ON DISCRIMINATING LETTERS FROM THEIR TRANSFORMATIONS. (JC)

ED003349

THE USE OF PROGRAMMED INSTRUCTION IN THE DEVELOPMENT OF RECOGNITION SKILLS

Cooperative Research Project No. 2100

JOANNA P. WILLIAMS
University of Pennsylvania
Philadelphia, Pennsylvania

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

THE USE OF PROGRAMMED INSTRUCTION IN THE DEVELOPMENT OF RECOGNITION SKILLS

Cooperative Research Project No. 2100

Joanna P. Williams

University of Pennsylvania
Philadelphia, Pennsylvania

1965

The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare.

ED 003349

TABLE OF CONTENTS

Page

Experiment 1: Retention of Introductory and Review Programs as a Function of Response Mode.................................. 1

Experiment 2: The Effectiveness of Constructed Response and Multiple Choice Programming Modes as a Function of Test Mode.................. 15

Experiment 3: Combining Response Modes in Programmed Instruction.............................. 35

Experiment 4: Recognition of Letter-like Forms in Appropriate Orientation.................. 53

TABLES AND FIGURES

Page

Experiment 1:

Table 1: Mean Number Correct on the Criterion Test... 7

Table 2: Analysis of Variance of Number Correct on Criterion Test.............................. 8

Table 3: Correlations between Aptitude and Performance Measures........................ 11

Experiment 2:

Table 1: Mean Number Correct on the Criterion Test... 21

Table 2: Mean Number Correct on the Constructed Response Criterion Test..................... 22

Table 3: Analysis of Variance of Number Correct on Constructed Response Test................ 23

Table 4: Mean Number Correct on the Multiple Choice Criterion Test...................... 25

Table 5: Analysis of Variance of Number Correct on Multiple Choice Test..................... 26

Table 6: Mean Gain (Posttest minus Pretest) on Each Criterion Test............................ 27

Experiment 3:

Table 1: Mean Number Correct on the Criterion Test... 41

Table 2: Mean Number Correct on the Constructed Response Criterion Test.................... 42

Page

Table 3: Analysis of Variance of Number Correct on Constructed Response Test............. 43

Table 4: Mean Gain (Posttest minus Pretest) on the Constructed Response Test............ 45

Table 5: Training time (in minutes)............... 47

Experiment 4:

Figure 1: The Standard Forms and their Transformations.......................... 59

Table 1: Proportion of Errors on Each Test......... 65

Table 2: Number of Errors as a function of Type of Training and Type of Transformation (Exp. 1)................................. 68

Table 3: Number of Errors as a function of Type of Training and Type of Transformation (Exp. 2)................................. 69

Table 4: Analysis of Variance of Number of Errors on the Three Tests (Exp. 1)............... 70

Table 5: Analysis of Variance of Number of Errors on the Three Tests (Exp. 2)........ 71

Table 6: Proportion of Errors on each Form and its Transformations........................ 73

References... 80

ACKNOWLEDGEMENTS

Ellen I. Levy served as Research Assistant on this Project. In this capacity, she was involved in all phases of the research, from running subjects and analyzing data to the actual planning of experiments. The author is very grateful to her for her invaluable assistance.

JoAnne K. Lesch and Patience Rockey were also involved in certain phases of the work reported here. Both contributed a great deal to the research, and their help is greatly appreciated.

Dr. I. Ezra Staples of the Philadelphia Board of Education very generously gave advice and guidance in selecting kindergartens and elementary schools for participation in the experiments. The author wishes to thank him, and also the principals, teachers, and pupils of the fourteen Philadelphia schools who co-operated.

Appreciation is also due to Coronet Instructional Films, Chicago, Illinois, for permission to reproduce a portion of their program GROUPING ANIMALS: WHAT IS A MAMMAL? in experimental versions.

The first two experiments have been published under

the same titles listed in this Report. Experiment 1 was published in the American Educational Research Journal, Volume 1, 1964, pages 211-218, by J. P. Williams and E. I. Levy. Experiment 2 was published in the Journal of Educational Psychology, Volume 56, 1965, pages 111-117, by J. P. Williams.

J.P.W.

RETENTION OF INTRODUCTORY AND REVIEW PROGRAMS AS A FUNCTION OF RESPONSE MODE

While there have been many studies of the effectiveness of active participation in programed instruction, most have presented data only from tests administered soon after the completion of the program (Evans, Glaser, and Homme, 1960; Keisler and McNeil, 1962; Williams, 1963). In general, performance is better when there is active participation, although often the comparisons between response modes have not reached conventional levels of significance. The little data available from tests given after some time has elapsed since training suggest that delayed testing might be more sensitive to training differences. For example, Krumboltz and Weisman (1962) found that subjects who had to construct their own responses performed significantly better on a test given two weeks after completion of the program than those who made covert responses and those who simply read the program; these differences had not appeared on the immediate test. Goldbeck and Campbell (1962) also found differences only on a delayed test; in this case, however, after 10 weeks, subjects in the group that read through the program performed significantly better than those who had had constructed-response training. One purpose of the present experiment was to collect further data on the effectiveness of constructed-response versus reading modes as a function of time of testing.

Another variable was also investigated in the present study: whether or not the content of the program had previously been studied. It has often been proposed that programed materials be used as review. For example, programed homework assignments reviewing material introduced by the teacher might be given, or a few hours' work on programed materials covering their previous work could be given to pupils returning to school after a long vacation. Although programed instruction does appear promising for this purpose, the recommendation has not been made on the basis of data. Holland and Porter (1961) investigated the value of review within a program (through repetition of items answered incorrectly during the first administration), but there have been no evaluations of an entire program used as a review lesson. Williams (1963) used a program designed as introductory material for review and found that it was at least as sensitive to the experimental conditions as when given as new content. Moreover, the subjects (college students) tended to prefer that the program be used as review. A second purpose of this study was to compare the effectiveness of the same program when used as introductory material and as review.

Programed instruction might indeed prove to be a very effective means of review, in light of the claim that it leads to more homogeneous performance levels. Several studies have found (e.g., Evans, Glaser, and Homme, 1960) that a group of subjects going through programed materials showed significantly less variability on a posttest than a group learning from a more conventional textbook or lecture approach. Williams (1963) found that a group given constructed-response training showed less variability than groups whose training was the same except in terms of response mode, and this indicated that under certain conditions, at least, active participation is one of the factors that influences the variance of the scores. Other investigators (Porter, 1959; Detambel and Stolurow, 1956) have found that correlations between aptitude and performance on programed materials are smaller than when the instruction is of some other less efficient type. Stolurow (1961) suggests that the decreased variability of posttest scores "may be the factor contributing to the drop in correlation between general ability measures and scores on programed materials." In the present study, the relationship of aptitude (Philadelphia Verbal Ability Test scores) and achievement was compared both on the posttest and in terms of amount of gain actually made as a result of going through the program.

Method

Subjects: Subjects were 96 sixth graders who had not been previously exposed in school to the material covered in the program. An additional 44 subjects were dropped: 19 for failure to complete all training and testing sessions, and 25 (randomly) to equate aptitudes and N's in the experimental groups.

Materials: The first 120 frames of the Coronet Program "Grouping of Animals: What is a Mammal?" were reproduced in booklet form. The first type of booklet consisted of the standard constructed-response format. Each frame appeared on a separate page with the correct response presented alone on the following page. The subject was required to write out the answer for every frame on a separate answer sheet. In the second type of booklet, each frame appeared as a complete statement, to be read as regular text material. Each frame appeared on a separate page; no answer pages or answer sheets were required.

In addition, four 22 x 28 inch posters were used as aids for those groups receiving preliminary instruction to illustrate the meaning of four technical terms introduced in the program. The technical term (e.g., "class"; "phylum Chordata") was written at the top of each poster above colored pictures of animals grouped appropriately.

An objective test was made up of 32 items, half of which required constructed responses and the other half, a choice between two alternatives. Half the items required the subject to utilize technical terms introduced in the program, and the other half were items to which the subjects could respond with familiar words in their standard vocabulary. The test was divided into two parts, each of which had to be completed independently. Within each part, each response term appeared only as a multiple-choice item or as a constructed-response item, so that there could be no cueing from one type of item to the other. Half the subjects were given part 1 first, and the other half, part 2 first. This test was used both as pretest and posttest.

Procedure: Four classrooms were used in the main design of the experiment. The experimenter gave two of these classes, selected at random, approximately 20 minutes of instruction on the material presented in the program. For these subjects, then, the program, given later, was essentially a review of previously learned material. The other two classes were given no such preliminary training. Within these two groups (Review Content and New Content), subjects were randomly assigned to one of the two training conditions (Constructed-Response mode or straight Reading mode). Ten days after the Review Content group had received its preliminary instruction, all subjects were administered the pretest, and four days later they were given the program. On each of two alternate days subjects went through 60 frames of the program. On the day following the completion of the program, half the subjects in each condition (randomly chosen) were given the posttest. The other subjects were given their posttest four weeks after the completion of the program. Sixteen minutes were allowed for the pretest (eight minutes for each half), and the same procedure was followed for the posttest. All subjects were able to finish within the time limit.

Results

Mean scores on the pretest and posttest as a function of type of content, response mode in training, and time of testing are presented in Table 1. Both sets of scores were normally distributed, and the variances of the two training modes did not differ on the posttest. There was a significant increase from pretest to posttest for all New Content subjects ($t=8.59$, $df=47$, $p<.001$), whereas a group of similar subjects who were given no experimental training showed no increase from pretest to posttest ($t=1.40$, $df=23$).

Analysis of variance (see Table 2) indicated that there were no significant differences on the pretest except as a function of type of content: subjects who were to get the program as review, having already been given the material through the preliminary teaching, scored higher than those who were to be introduced to the material by the program.

Table 1

Mean Number Correct on the Criterion Test

	Pretest		Posttest	
	New Content	Review Content	New Content	Review Content
Immediate test				
CR training	13.16	18.75	20.08	23.00
R training	14.75	18.91	20.42	21.08
Delayed test				
CR training	13.00	19.25	18.75	22.83
R training	13.58	19.83	16.83	21.50

Table 2

Analysis of Variance of Number Correct on Criterion Test

Source	df	Pretest MS	Pretest F	Posttest MS	Posttest F	Difference Score MS	Difference Score F
Type of Content (A)	1	742.60	45.26**	160.16	10.56**	228.16	22.68**
Response Mode (B)	1	12.76	<1	70.04	4.62*	155.04	14.96**
Time of Testing (C)	1	0.01	<1	66.66	4.39*	77.04	7.43**
A x B	1	3.01	<1	0.17	<1	7.05	<1
A x C	1	11.34	<1	77.05	5.08*	35.05	3.38
B x C	1	0.51	<1	0.17	<1	2.67	<1
A x B x C	1	3.01	<1	1.04	<1	1.49	<1
Error	88	16.41		15.17		10.36	

*$p < .05$

**$p < .01$

On the posttest all three variables showed significant effects: the Review Content group was superior to the New Content group, the Constructed-Response group scored higher than the Reading group, and scores on the Immediate posttest were higher than those on the Delayed test. However, these differences must be interpreted with caution because the interaction between type of content and time of testing was significant. Analysis of the simple effects indicated that for those having the program as review, there was no difference between the immediate and delayed tests, whereas for those taking the program as introductory material, performance on the delayed test was significantly below performance on the immediate test.

The difference between pretest and posttest scores was taken as a measure of the amount gained from the experimental training (Goldbeck and Campbell, 1962). The analysis of variance of these scores is also presented in Table 2 . The results were for the most part similar to those of the analysis of posttest performance: constructed-response training was significantly better than reading training, and performance on the immediate test was significantly higher than that on the delayed test. None of the interactions reached the .05 level of significance. As before, there was no difference between immediate and delayed testing for the Review Content group, whereas there was a reliable drop from the immediate testing to the delayed testing in the New Content group. The New Content group showed greater gains than the Review Content group, in contrast to the results shown on the posttest. (This comparison, however, is rather meaningless because the pretest scores of the two groups were significantly different, of course, and there is no way of establishing the equality of the units at different levels of the test.)

Correlations between aptitude and each of the performance measures are presented in Table 3. As expected, there was a positive correlation between aptitude and achievement, whatever measure was used. However, the correlations between aptitude and posttest score were larger (7 of the 8 were significant at the .05 level, and the other reached the .10 level), whereas none of the correlations between aptitude and difference score, representing actual gains from the program, reached conventional levels of significance, and in fact, two were negative. There were no differences in the correlations as a function of type of content, time of testing, or response mode.

An over-all estimate of the correlations for each of the performance measures was made: for the pretest, it was .56, ($p < .01$); for the posttest, .69, ($p < .01$); for the difference scores, .15, not significant. The latter was significantly different from the other two.

Table 3

Correlations between Aptitude and the Performance Measures

	Pretest		Posttest		Difference Score	
	New Content	Review Content	New Content	Review Content	New Content	Review Content
Immediate test						
CR training	.60*	.73**	.69**	.80**	.21	.11
R training	.40	.88**	.58*	.58*	.31	-.54
Delayed test						
CR training	.05	.36	.70**	.71**	.49	.26
R training	.70**	.36	.85**	.50	-.15	.25

*$p<.05$

**$p<.01$

N=12 in each case

Discussion

The finding that performance was a function of active participation replicated earlier findings. The fact that the difference between Constructed-Response and Reading modes was seen not only when the program presented an introduction to content but also when the program served as review, and not only on the immediate test but also on the delayed test, helps establish the generality of the finding.

The data of this study do not agree with those of Krumboltz and Weisman (1962) and of Goldbeck and Campbell (1962), who found differences between response modes only on a delayed test. In those experiments all subjects were given both posttests, and perhaps the significant differences on the delayed test were in some way due to the effects of the immediate test. Both the Krumboltz and Weisman (1962) experiment and the present one found better performance in the constructed-response group on the delayed posttest, as expected from previous studies involving the active participation variable.

There were no differences in the variance of the posttest scores as a function of response mode in training, contrary to previous findings (Williams, 1963). Thus, in this study, active participation does not seem to be an important factor in modifying the range of individual differences in performance. Elementary-school students were used in the current study whereas Williams (1963) used college students, and this difference may account for the discrepant findings.

Previous work has indicated that while there may be a substantial correlation between aptitude and achievement on a posttest after programed instruction, the correlations decrease markedly when actual gains are measured. In fact, some investigators have claimed that these correlations are essentially zero (Detambel and Stolurow, 1956; Porter, 1959). The present study corroborated these findings. There was a rather large and reliable correlation between aptitude and performance on the posttest, but the correlation between aptitude and difference score was small and not significantly different from zero. While it would be inappropriate to conclude that there is _no_ correlation between aptitude and gains on the **basis** of the current rather small sample (and in most cases previous studies utilized Ns that were considerably smaller than the present one), it seems clear that the relationship in question is indeed smaller than that between aptitude and posttest. These differences in the correlations cannot be attributed to an attenuation of the range in the case of the difference scores; the standard deviations of the two performance measures did not differ.

Summary

Sixth-grade pupils worked through a linear program in one of two versions: (a) standard constructed-response format, and (b) as straight text, requiring no active participation. The constructed-response group performed better on an objective criterion test. Subjects were tested either on the day following completion of training or one month later; scores on the immediate test were higher than those on the delayed test. For half the subjects, the program presented a review of previously learned material, and these subjects performed better than those for whom the material was new. There was a significant interaction between type of content and time of testing. An analysis of the actual gains made from the program showed rather similar results. There were significant correlations between aptitude and pretest score and between aptitude and posttest score, but the correlation between aptitude and gain score (difference between pretest and posttest) was not significant.

THE EFFECTIVENESS OF CONSTRUCTED RESPONSE AND MULTIPLE CHOICE PROGRAMMING MODES AS A FUNCTION OF TEST MODE

Recent studies have suggested that there is no simple answer to the question of what is the most appropriate mode of response in programmed instruction. For example, Williams (1963) found that the relative effectiveness of two training modes may depend on the type of item used in the criterion test. In that experiment, college students worked through a program which reviewed previously learned material, and a constructed-response criterion test was administered. It was found that constructed-response training was superior to multiple-choice training on test items requiring the use of novel or technical terminology which had been introduced by the program, but not on items requiring general, familiar vocabulary. The present experiment was designed to extend this finding to a situation (1) where the content of the program was novel, and not review; and (2) where the subjects were elementary school pupils. In addition, the present experiment utilized not only a constructed-response criterion test but also a multiple-choice test, in order to determine whether the latter would also be sensitive enough to show the predicted interaction between response mode in training and type of test item.

In the previous study, only posttest scores were available; here, the criterion test was also administered as a pretest, so that the relationship between the variables could also be investigated in terms of the actual gains made as a result of having gone through the program.

A second purpose of the experiment was to examine the relationship between performance on programmed materials and aptitude. It is reasonably well established at the present time that aptitude and achievement, as measured by posttest performance, are positively related when programmed materials are used (Gleason, 1964; Lambert, Miller and Wiley, 1962; Williams, 1963). However, in most cases, these correlations may have been due to differences which were present before the program was administered; for it is to be expected (and indeed Gleason (1964) has found) that pretest performance is also a function of aptitude. It may be, as has been suggested, for example, by Ferster and Sapon (1958) and by Stolurow (1961), that when actual gains are assessed, there is little or no correlation between aptitude and performance. In the present experiment, the relationship of achievement and aptitude was examined both in terms of posttest performance and in terms of amount of gain from pretest to posttest.

Method

Subjects: Subjects were 108 sixth-grade pupils who had not previously been exposed in school to the material taught in the program. Half the subjects were taken from high aptitude classrooms (mean I.Q. about 125) and half, from average aptitude classrooms (mean I.Q. about 100). Aptitude scores (Philadelphia Verbal Ability Test) were normally distributed, and the variances of the scores at the two aptitude levels did not differ. The subjects were run in two replications, 60 in the first and 48 in the second. Twenty-five additional subjects were dropped, 19 for failure to complete all training and testing sessions, and 6 (randomly) in order to equate N in each group.

Materials: The first 120 frames of the Coronet Program "Grouping Animals: What is a Mammal?" (1962) were reproduced in booklet form. Each frame appeared on a separate page, with the correct response presented alone on the following page. The program was reproduced in two forms: (a) Constructed Response (CR): standard format, in which the subject was required to write out answers for every frame; and (b) Multiple Choice (MC): under each item, presented in the standard way, two response alternatives were printed, and the subject was instructed to underline his choice. All subjects recorded their responses on separate answer sheets.

An objective test was made up of 16 items which required constructed responses. Half the items required the subject to utilize technical terms introduced in the program (such as vertebrae and phylum), and the other half were items to which the subjects could respond with familiar words in his standard vocabulary. An additional 16 items, which required a choice between two alternative answers, were also given; half demanded technical terms as responses, and half were of the familiar vocabulary type. These were analyzed as two separate tests, but they were presented to the subject as one test, in which items of both kinds -- constructed-response and multiple-choice -- were mixed together.

Items were balanced so that every technical term which was used as a response appeared both on the CR and on the MC tests. In addition, the test was divided into two parts, each of which had to be completed independently. In each half, each response term appeared either as a MC or as a CR item, so that there could be no cueing from one type to the other. Half the subjects were given Part 1 first and then Part 2, and for the other subjects, the order was reversed.

Procedure: Subjects were randomly assigned to one of the two training conditions (CR or MC). All subjects were given the objective test as a pretest, and sixteen minutes were allowed (eight minutes for each half). On each of three successive days of training, starting four days after the pretest, subjects worked through 40 frames of the program. The time taken by each subject to work through each day's assignment was recorded. On the day following completion of the program, the subjects were again given the objective test with the same time limits as in the pretest. All subjects finished within the time allowed.

Before the experimental training, subjects were told that this was an experiment in programmed instruction and that their class had been specially chosen for participation. It was pointed out that their test scores would not count as part of their school grades, but that they should try to work as well as possible. This type of introduction was chosen on the basis of advice from classroom teachers, and it was felt that it provided good motivation.

Results

Because there were no differences between the two replications in the mean number of correct responses on the pretest as a function of response mode (High Aptitude (HA): $F<1$, $df=1/50$; Low Aptitude (LA): $F=1.23$, $df=1/50$)

or on the posttest (HA: F=1.67, df=1/50; LA: F=1.73, df=1/50), or in aptitude (HA: F=1.62, df=1/50; LA: F=1.02, df=1/50), the two replications were combined for analysis.

Means and standard deviations of the performance measures are presented in Table 1. Both pretest and posttest scores were normally distributed, and contrary to previous findings (Williams, 1963), the variances of the two training modes did not differ on the posttest. There was a significant increase from pretest to posttest on both the constructed-response (CR) test and the multiple-choice (MC) test. A group of subjects who were given no experimental training showed no increase from pretest to posttest.

Table 2 presents the mean number correct on the constructed-response criterion test as a function of training mode, aptitude, and type of test item. As indicated by analysis of variance (summarized in Table 3), there were no differences on the pretest as a function of response mode in training. HA subjects scored significantly higher than LA subjects, and scores on general vocabulary items were higher than those for technical items. There were no interactions. On the posttest, as on the pretest, aptitude (HA>LA) and type of test item (general>technical) showed significant effects. In addition, subjects who had CR training scored significantly higher than those who had MC training. As predicted, there was a significant interaction

Table 1

Mean Number Correct on the Criterion Test

Test Mode	Training Mode	Pretest Mean	Pretest S.D.	Posttest Mean	Posttest S.D.	t
CR	CR	4.00	2.22	9.68	3.29	8.67*
	MC	3.48	2.16	7.70	3.35	7.81*
	None	4.20	2.37	5.15	2.78	1.10+
MC	CR	10.76	2.05	13.67	1.45	8.56*
	MC	10.31	2.33	13.68	1.54	8.87*
	None	10.35	2.32	11.05	1.70	1.14+

*df=106, p<.001

+df=38, not significant

Table 2

Mean Number Correct on the Constructed Response Criterion Test

Training mode	Aptitude	Pretest General items	Pretest Technical items	Posttest General items	Posttest Technical items
CR	High	3.59	1.15	6.22	5.41
	Low	2.56	0.70	4.52	3.22
MC	High	2.89	1.22	5.30	4.18
	Low	2.26	0.59	4.07	1.85

Table 3

Analysis of Variance of Number Correct on Constructed Response Test

Source	df	Pretest MS	Pretest F	Posttest MS	Posttest F	Gain MS	Gain F
Training mode (A)	1	3.63	1.27	53.01	13.88**	26.04	7.42**
Aptitude (B)	1	25.35	12.25**	187.04	48.96**	77.04	21.95**
A x B	1	0.17	<1	0.37	<1	0.11	<1
Error (b)	104	2.07		3.82		3.51	
Type of test item (C)	1	196.46	148.83**	100.04	104.21**	15.04	16.00**
A x C	1	3.13	2.37	5.04	5.25*	15.04	16.00**
B x C	1	1.19	<1	8.56	8.92**	16.11	17.13**
A x B x C	1	1.18	<1	1.34	1.40	0.02	<1
Error (w)	104	1.32		0.96		0.94	

*p<.05

**p<.01

between training mode and type of test item, and analysis of the simple effects indicated that the differences between the two training modes appeared on technical terms, but not on general terms. There was also an interaction between aptitude and type of test item: the difference between scores on technical items and general items was larger for the LA group than for the HA group.

Table 4 and Table 5 present the data from the multiple-choice test. On the pretest, HA subjects scored significantly higher than did LA subjects, and there was a tendency for scores to be higher for technical terms than for general terms ($p<.10$). Training mode was not a significant effect, and there were no interactions. On the posttest, aptitude (HA>LA) and type of test item (technical> general) were significant effects. There was no significant difference between the training modes, and there were no interactions. It should be noted that two of the three within-subjects error terms exceed the corresponding between-subjects error terms. Although the data are not available to make a satisfactory assessment, the possibility that the MC test was not highly reliable should be considered.

The difference between pretest and posttest scores was taken as a measure of the amount gained from the experimental training, and the results are presented in Table 6.

Table 4

Mean Number Correct on the Multiple Choice Criterion Test

Training mode	Aptitude	Pretest General items	 Technical items	Posttest General items	 Technical items
CR	High	5.52	6.26	6.89	7.40
	Low	4.78	4.96	6.15	6.89
MC	High	5.44	5.78	6.89	7.48
	Low	4.59	4.81	6.04	6.96

Table 5

Analysis of Variance of Number Correct on Multiple Choice Test

Source	df	Pretest		Posttest		Gain	
		MS	F	MS	F	MS	F
Training mode (A)	1	2.66	2.42	0.01	<1	1.67	1.04
Aptitude (B)	1	50.07	45.52**	23.34	25.37**	10.23	6.39*
A x B	1	0.17	<1	0.04	<1	12.04	7.53**
Error (b)	104	1.10		0.92		1.60	
Type of test item (C)	1	7.40	3.70	26.04	35.19**	0.01	<1
A x C	1	0.47	<1	0.22	<1	1.34	<1
B x C	1	1.51	<1	1.04	1.41	0.04	<1
A x B x C	1	0.66	<1	0.05	<1	7.30	2.92
Error (w)	104	2.00		0.74		2.50	

*$p < .05$

**$p < .01$

Table 6

Mean Gain (Posttest minus Pretest) on Each Criterion Test

Training mode	Aptitude	CR Test General items	CR Test Technical items	MC Test General items	MC Test Technical items
CR	High	2.63	4.22	1.37	1.18
	Low	1.92	2.44	1.37	1.92
MC	High	2.41	2.96	1.44	1.70
	Low	1.81	1.26	1.52	2.15

Analysis of the gains on the CR test show results very similar to those of the analysis of CR posttest performance: and thus the analysis of variance summary is not presented. CR subjects gained more than those who had MC training, and HA subjects gained more than did LA subjects. There was no interaction between these two variables. There was more gain on items requiring technical terminology than on general items. The predicted interaction between training mode and type of test item was significant. Analysis of the simple effects showed that the difference between gains made by the two training modes appeared only on the technical items; on general terms, the two training modes gained equally. The interaction between aptitude and type of test item was also significant, as in the posttest analysis: for HA subjects, there was more gain on technical terms than on general terms, while for LA subjects, the gains on the two types of item were equal.

Table 6 also shows the mean gains on the multiple choice test. LA subjects gained significantly more than did HA subjects. No other main effect was significant. There was one significant interaction, between training mode and aptitude, such that for HA subjects, the two training modes were equal, but for LA subjects, the CR training group gained more than did the MC training group.

CR training took significantly longer (M=47.5 min.)

than MC training (M=35.8 min.) (F=45.60, df=1/104, $p < .001$). There was no significant difference between the two aptitude groups (F=1.50, df=1/104) and the two variables did not interact (F=2.35, df=1/104). An efficiency score was calculated for each subject by dividing his posttest score by his training time, giving a measure of amount learned per unit of learning time (Goldbeck and Campbell, 1962). MC training was significantly more efficient than CR training (F=9.82, df=1/104, $p < .01$); HA subjects were more efficient than LA subjects (F=26.34, df=1/104, $p < .001$); and the two variables did not interact ($F < 1$). An efficiency measure based on actual gains (difference score divided by training time) showed a slightly different pattern, however. As before, HA subjects were more efficient than LA subjects (F=8.00, df=1/104, $p < .01$), but there was no effect of training mode ($F < 1$). There was an interaction (F=4.00, df=1/104, $p < .05$) such that for HA subjects there was no difference between the training modes, but for LA subjects, MC training was significantly more efficient.

Discussion

The predicted interaction between training mode and type of test item was significant both on the CR posttest and also when the CR difference scores were analyzed. This finding supports the initial hypothesis of the experiment, namely, that CR training is more effective than MC training when complex, technical terminology introduced by the program is required, as compared to when general and familiar vocabulary is required.

The fact that the main effect of training mode was significant on the posttest was unexpected, for to the writer's knowledge, no previous study utilizing a typical continuous-discourse program has shown CR and MC modes to differ significantly. The only previous study showing a significant overall difference between these two modes was that of Fry (1960). In that experiment, the test items required the use of newly-acquired Spanish vocabulary; these clearly would fall into the category of technical terminology. In Williams' (1963) experiment, where the overall comparison of CR and MC training did not reach significance, but a comparison involving only technical terms did, the criterion test was taken from material designed as part of an actual instructional program. Less than one-third of the items required technical terms as responses, and it is likely that tests used with continuous-discourse programs generally include a relatively low ratio of items which require difficult, complex, "technical" responses. The present data suggest that CR training is superior to MC training to the extent to which the test demands use of such terms.

There was no effect of training mode in terms of the multiple-choice test, however, nor was there any interaction between training mode and type of test item, indicating that the more difficult constructed-response test reflected differences in the training procedures to which the relatively less demanding multiple-choice test was not sensitive.

The results show clearly that there is a relationship between aptitude and posttest score. Indeed, if we consider posttest score as the index of achievement--and this is how achievement is usually assessed in the classroom--the conclusion is inescapable that achievement on programmed materials is a function of aptitude. Moreover, there is a relationship between aptitude and pretest score, as Gleason (1964) found, indicating that there were differences even before the experimental training. In contrast, however, to studies which showed little or no relationship between aptitude and the amount actually gained from programmed materials (e.g., Ferster and Sapon, 1958; Porter, 1959; Williams and Levy, 1964), the present results clearly indicate such a relationship when the more sensitive CR test is analyzed. The posttest differences as a function of aptitude are thus partly attributable to differences present prior to the experiment and partly to the effects of the experimental training.

On the multiple-choice test, however, LA subjects showed higher gains than did HA subjects. This anomalous result is probably due to the ceiling imposed by the test: some of the HA subjects scored the maximum 16 points on the MC posttest, and it is most likely that they would have shown more improvement if the test had allowed it. Indeed, of those subjects who achieved a maximum score, 86% were in the HA group. In any event, HA subjects were still significantly superior to LA subjects on both posttests, and it is evident that differences in proficiency already present at the start of the experiment were not removed by administering the program, as expected by those who claim that with effective methods of teaching, such as programmed instruction, aptitude differences tend to lose their value in predicting achievement.

It should be noted that the fact that there was no effect of training mode on the MC test is not due to a ceiling effect. If such were the case, it would be expected that more CR training subjects (or MC training subjects) should score at the maximum, suggesting that that group might have made further gains which the test could not show. However, the subjects who scored at the maximum were almost equally divided between CR and MC training subjects (43% and 57%, respectively), so that the likelihood of achieving the maximum score was not different for the two groups.

While the CR training group scored significantly higher than the MC training group, the size of the difference was relatively small. It might be felt that for practical purposes such small increments are not worth the greater amount of time taken in training. However, the experimental program was a relatively short one, and the small effects seen here would be expected to cumulate. Thus, in any actual instructional sequence, or over the course of several instructional sequences, the differences between these two training modes would be large enough to be of some consequence.

It has often been suggested that several different response modes should be incorporated into a sequence of frames for the purpose of introducing variety into the program. The present data suggest that the effectiveness of such variation would be increased if the constructed-response mode were used on frames requiring novel, complex, or technical terminology, and other modes were introduced elsewhere.

Summary

Constructed response and multiple choice versions of a program were administered to sixth-grade pupils. Per-

formance on a constructed-response test was superior for the constructed-response training mode, and high-aptitude subjects were significantly superior to low-aptitude subjects. The predicted interaction between training mode and type of test item (those requiring novel, "technical" terminology as responses vs. those requiring general, familiar vocabulary) was significant. These effects were seen when posttest performance was analyzed and also when actual gains (posttest score minus pretest score) were assessed. Performance on a multiple-choice test was also a function of aptitude, but this test was not as sensitive to differences in training mode.

COMBINING RESPONSE MODES IN PROGRAMMED INSTRUCTION

An instructional method is evaluated according to two general criteria. First, the effectiveness of the method must be assessed. Does it teach what it is supposed to teach? Or, put another way, does performance on the criterion test reach the desired level? Secondly, the efficiency of the method must be taken into account, for methods that require unduly large amounts of training are not practical.

Constructed-response training, in which the student is required to write a response to every frame, has compared favorably in effectiveness with all other modes typically investigated (e.g., Holland, 1960; Krumboltz and Weisman, 1962; Williams, 1963). However, this mode is relatively inefficient: less is learned in a given amount of time than with any other training mode (Goldbeck and Campbell, 1962; Williams, 1963).

One variable that determines the relative effectiveness of different training modes is the type of item used on the criterion test. Constructed-response training, for example, is superior to multiple-choice training on test items that use novel or technical terminology, but not on items that use general, familiar vocabulary. This has been demonstrated with both new and review material, and in both elementary pupils and college students (Williams, 1963; Williams, 1965). As usual, however, constructed-response training required more time, and the overall efficiency of that mode was lower.

These findings suggest one way in which it might be possible to increase the efficiency of a program without diminishing its effectiveness: use constructed-response training on frames requiring technical terminology and multiple-choice training on the other items. How would such a program compare in (a) effectiveness and (b) efficiency with a program composed entirely of constructed-response frames and one which is entirely multiple-choice? The present study compares programs of these three types.

In order to evaluate such a combined-mode program properly, one must take account of possible interfering effects introduced by requiring the concurrent use of two response modes. Since the effects of introducing a mixture of training modes are unknown (Taber, Glaser, and Schaefer, 1965), a fourth program was designed, which also used the same two training modes, but in which there was no correlation between training mode and type of item. That is, frames requiring technical responses and those requiring general responses were assigned randomly to constructed-response presentation or to multiple-choice presentation.

Method

Subjects: Subjects were 120 sixth graders who had not been previously exposed in school to the material covered in the program. An additional 14 subjects were dropped: 11 for failure to complete all training and testing sessions, and three to equate pretest scores and Ns in the experimental groups.

Materials: The first 120 frames of the Coronet Program "Grouping Animals: What is a Mammal?" were reproduced in booklet form. Eighty-eight of these frames, 44 requiring technical terms as the response (e.g., vertebrae and phylum) and 44 requiring familiar vocabulary, were specified for experimental variation. The remaining 32 frames, which were distributed randomly throughout the booklet, appeared as complete statements, one per page, to be read as regular text material. No overt response was required. These frames were presented in this manner in order to balance the number of technical and general items to be manipulated in the four training groups.

In Group A, all 88 frames consisted of the standard constructed-response format. Each frame appeared on a separate page with the correct response presented alone on a following page. The subject was required to write out the answer for every frame on a separate answer sheet. In Group B, all 88 frames consisted of the multiple-choice format. Under each item, presented in the same way as in Group A, two alternative responses were printed. The subject was required to underline his choice, and he found the correct response on the following page. In Group C, those 44 frames which required technical terminology as the response were presented in the constructed-response format; the other 44 frames, which required general vocabulary, were presented in the multiple-choice mode. In Group D, 44 frames, randomly selected from the 88, were presented in the constructed-response format, and the other 44 appeared as multiple-choice items.

All subjects recorded their responses on a separate answer sheet, which contained blanks for the constructed responses and the two alternative choices in the multiple-choice items. The frames that were presented as complete statements and required no response were indicated by a row of _X_'s by the appropriate number.

An objective test was made up of 16 items that required constructed responses. Half the items required the subject to utilize technical terms introduced in the program (such as _vertebrae_ and _phylum_), and the other half were items to which the subject could respond with familiar words in his standard vocabulary. An additional 16 items, which required a choice between two alternative answers, were also given; half demanded technical terms as responses, and half were of the familiar vocabulary type. These were analyzed as two separate tests, but they were presented to the subject as one test, in which items of both kinds--constructed-response and multiple-choice--were mixed together.

Items were balanced so that every technical term which was used as a response appeared both on the CR and on the MC tests. In addition, the test was divided into two parts, each of which had to be completed independently. In each half, each response term appeared either as a MC or as a CR item, so that there could be no cueing from one type to another. Half the subjects were given Part 1 first and then Part 2, and for the other subjects, the order was reversed.

Procedure: Subjects were randomly assigned to one of the four training conditions. All subjects were given the objective test as a pretest, and sixteen minutes were allowed (eight minutes for each half). On each of two successive days of training, starting three days after the pretest, subjects worked through 60 frames of the program. The time taken by each subject to work through each day's assignment was recorded. On the day following the completion of the program, the subjects were again given the objective test with the same time limits as in the pretest. All subjects finished the test in the time allowed.

Before the experiment began, subjects were told that this was an experiment in programmed instruction and that their class had been specially chosen for participation. It was pointed out that their test scores would not count as part of their school grades, but that they should try to do as well as possible.

Results

Means and standard deviations of the performance measures on the constructed-response criterion test are presented in Table 1. Both pretest and posttest scores were normally distributed. There was a significant increase from pretest to posttest. An untrained control group was considered unnecessary since previous work (Williams, 1965) indicated no difference between pretest and posttest following similar manipulations.

Table 2 presents the mean number of correct responses on the constructed-response criterion test as a function of training mode and type of test item. As indicated by analysis of variance (summarized in Table 3), there were no differences on the pretest as a function of response mode in training. Scores on general vocabulary items were higher than those for technical items. There were no significant interactions. On the posttest, as on the pretest, type of test item (general>technical) showed a significant effect. In addition, the differences due to

Table 1

Mean Number Correct on the Constructed-Response Criterion Test

Training Mode	Pretest		Posttest		t
	M	SD	M	SD	
A (All Constructed Response)	2.97	2.12	8.00	3.90	6.06*
B (All Multiple Choice)	2.57	2.03	6.10	3.21	4.97*
C (Combined, Correlated)	2.67	1.88	7.00	3.72	5.55*
D (Combined, Uncorrelated)	2.67	1.65	6.10	3.05	5.28*

df=58, p<.01

Table 2

Mean Number Correct on the Constructed-Response Test

Training Mode	Pretest		Posttest	
	Technical Terms	General Terms	Technical Terms	General Terms
A(All Constructed Response)	0.73	2.23	3.30	4.70
B(All Multiple Choice)	0.60	1.97	2.07	4.03
C(Combined, Correlated)	0.50	2.17	2.87	4.13
D(Combined, Uncorrelated)	0.57	2.10	2.40	3.70

Table 3

Analysis of Variance of Number Correct on Constructed-Response Test

		Pretest		Posttest		Gain	
Source	df	MS	F	MS	F	MS	F
Training Mode (A)	3	0.45	<1	12.30	2.02*	8.45	2.48*
Error (b)	116	1.86	--	6.09	--	3.41	--
Type of Item (B)	1	138.01	105.34**	132.02	84.63**	0.07	<1
A x B	3	0.22	<1	1.60	1.03	2.89	1.59
Error (w)	116	1.31	--	1.56	--	1.82	--

*$p<.10$

**$p<.01$

training mode tended toward significance. The two variables did not interact.

Since a priori predictions had been made about the training mode variable, specific comparisons were made. They indicated that Group A (all CR) was not significantly different from Group C (combined modes, correlated with item type) ($F=1.23$, $df=1/116$), but that Groups A and C together were significantly higher than Groups B (all MC) and D (combined modes but uncorrelated with item type) ($F=4.71$, $df=1/116$, $p<.05$). The latter two groups did not differ ($F<1$).

The difference between pretest and posttest scores was taken as a measure of the amount gained from the experimental training, and the results are presented in Table 4. Analysis of the gains on the CR test show results very similar to those of the analysis of the CR posttest. (The analysis of the variance summary is presented in Table 3.) Again, training mode reached the .10 level of significance. The two types of item were not significantly different, and there was no interaction.

Specific comparisons on these gain scores showed the same pattern as did those on the posttest. Group A was not reliably higher than Group C ($F=1.08$, $df=1/116$). Groups A and C scored significantly higher than Groups B and D

Table 4

Mean Gain (Posttest minus Pretest)
on the Constructed-Response Test

Training Mode	Technical Terms	General Terms
A (All Constructed Response)	2.57	2.47
B (All Multiple Choice)	1.47	2.07
C (Combined, Correlated)	2.37	1.97
D (Combined, Uncorrelated)	1.83	1.60

($F=6.33$, $df=1/116$, $p<.05$). The latter two groups did not differ ($F<1$).

Means and standard deviations of the amount of time taken in training are presented in Table 5. There were no differences among the experimental groups ($F=1.82$, $df=3/116$). An efficiency score was calculated for each subject by dividing his posttest score by his training time, giving a measure of amount learned per unit of training time (Goldbeck and Campbell, 1962). There were no differences among the groups on this measure ($F=1.16$, $df=3/116$). An efficiency measure based on actual gains (difference score divided by training time) also failed to show differences ($F=1.60$, $df=3/116$).

There was a significant improvement from pretest to posttest on the multiple choice criterion test ($p<.01$ for each group), but, as in a previous study (Williams, 1965), this test was not sensitive to differences in training mode ($F<1$, $df=3/116$, on both posttest and gain scores).

Discussion

In terms of the effectiveness criterion, i.e., achievement scores, the results were as expected. The superior posttest scores and gain scores of the constructed-response training group (A) as compared with the multiple-

Table 5

Training Time (in minutes)

Training Mode	Mean	S.D.
A (All Constructed Response)	39.4	7.5
B (All Multiple Choice)	38.6	9.6
C (Combined, Correlated)	43.8	10.2
D (Combined, Uncorrelated)	40.2	9.2

choice training group (B), previously demonstrated by Fry (1960) and Williams (1963, 1965) on constructed-response tests, also were found in the present study, where straight reading frames were incorporated into the programs. Since this comparison reached significance, the present experimental situation was considered suitable for assessing the relative effectiveness of the combined modes.

The important comparison is that between Groups C and B. As predicted, the program that combined constructed-response training on technical items and multiple-choice training on other items (Group C) was more effective than the wholly multiple-choice program (Group B), indicating that a combined mode can indeed boost performance. Moreover, it is clear that it was the correlation between CR training and technical terminology that accounted for Group C's superiority, and not the mixed format. Group D, which contained the same mixed format and the same amount of CR training but uncorrelated with item type, was significantly below Group C, and in fact was no better than Group B.

Many people have criticized programmed instruction on the grounds that it becomes monotonous and unstimulating (e.g., Fitzgerald, 1962). It is claimed that the routine, highly limited format tends to restrict independent thinking

(Doty and Doty, 1964). Such concerns have led to the reasonable suggestion (Taber, Glaser, and Schaefer, 1965) that there be some variation in the format of the program.

The present data suggest, however, that introducing a combination of modes into a program may actually lessen its effectiveness. It may be that the use of a combination of several training modes is confusing. Support for this notion comes from the fact that the performance of Group C, while not significantly different from that of Group A, was somewhat lower.

In addition, Group D, in which some of the frames were presented in the CR mode (some of which were on frames requiring technical terminology), was no better than Group B (all MC). Possibly the disadvantage of having no CR training in Group B was balanced by the disadvantage in Group D of having a mixed format. Moreover, when scores on technical and general items were examined separately, Groups B and D differed in effectiveness. Group B scored higher than D on general items. On these items, MC training is as effective as CR training, so that Group D had no advantage over Group B, whereas Group D had the disadvantage of the variety of modes. On the other hand, the fact that Group D was better than Group B on technical items indicates that the importance of CR training on items requiring technical terminology as responses is so great that even the partial matching of CR training and technical items in Group D was enough to raise its performance level above that of B.

Thus a mixed format seems to interfere with learning and not to enhance it. Unless methods can be worked out whereby the introduction of several response modes in training does not lead to confusion and interference, it would probably be wise to use only one response mode in training. Other methods of incorporating variety into a program, such as interspersing "test-type" frames that have no formal confirmation (Hershberger's (1964) self-evaluation response items), might prove more useful.

These comments are, of course, based only on trends. It is felt, however, that while no conclusions on these points can be drawn from the present study, further investigation is warranted.

The data are not conclusive with respect to the second criterion, efficiency. Contrary to expectation, there were no significant differences among the groups in the amount of time taken to go through the program. The addition of the large number of reading frames, which all groups could go through at a rapid pace, probably minimized overall training time differences. Interestingly, although the differences were not large enough to be significant, the two groups trained with combined modes tended to take longer than the other groups, lending further support for the idea that variety tends to be confusing.

Since there were no differences in training time, one would expect that the differences in achievement level would be reflected in the efficiency measures. This was not the case, however, probably because of the fact that there was so much variance in the time scores and such small differences on the posttest. In any event, since Groups A (all CR) and B (all MC) did not show the expected difference in efficiency, the present conditions did not afford an adequate situation in which to test the hypothesis that a combined mode where mode is correlated with item-type would be more efficient than an all-constructed-response program.

Summary

Sixth-grade pupils worked through a linear program in one of four versions: (A) standard constructed-response format; (B) standard multiple-choice format; (C) a combination format in which the constructed-response mode was used on frames that required technical terminology as responses and the multiple-choice mode on items where

2

general vocabulary was required; and (D) a combination format in which response-mode was uncorrelated with item-type. Groups A and C were significantly better than Groups B and D on a constructed-response posttest and on gain scores. The groups did not vary in efficiency. The findings suggested that using a combination of modes in training may lead to confusion and may reduce the effectiveness of the program.

RECOGNITION OF LETTER-LIKE FORMS IN APPROPRIATE ORIENTATION

To differentiate and recognize letters is one of the primary steps in learning to read, and this task is often a major source of difficulty in reading instruction (Vernon, 1957). Not only are certain letters of the alphabet quite similar in form to others, but in some cases they are merely rotations and reversals of others, e.g., b, d, and q, p. These letters, in fact, produce the most confusion and difficulty (Smith, 1928; Gibson, 1962; Popp, 1964).

Many practice materials have been designed to handle this problem. Some stress discrimination training (matching-to-sample and sorting pictures, symbols, or actual letters), whereas others emphasize reproduction training of some type (tracing or copying). Reading methods themselves vary as to the emphasis placed on instruction in writing. Most methods incorporate writing instruction into regular reading lessons, and some even stress it (Fernald, 1943; Spalding and Spalding, 1957); yet at least one recently developed method (Moore, 1963) puts off writing until late stages of training.

Evidence from the psychological literature as to the relative effectiveness of discrimination training and reproduction training is meager. It is often held that one should directly train the behavior which is to be tested. According to this principle, training in discrimination should be given if the criterion is some type of recognition skill. On the other hand, it might be argued that increasing the degree of active participation by requiring the subject to reproduce the material in some manner might lead to superior performance even on a recognition task. Reproduction training, of course, is a form of constructed-response training, which in other contexts has been demonstrated to be distinctly superior to multiple-choice (discrimination) training (Fry, 1960; Williams, 1963, 1965).

Analysis of what underlies the development of discrimination may be helpful. Pick (1965) has presented evidence for Gibson's (1962) hypothesis that improvement of visual discrimination depends on learning the distinctive features of the forms to be discriminated, i.e., those dimensions of difference that distinguish the stimuli. Gibson (1963, 1965) has undertaken the task of precise specification of the critical features of letters. Hopefully, when this is accomplished, those features that are essential for discrimination will be determined and training methods emphasizing those features developed.

Until this large-scale task is completed, however, perhaps it might be worthwhile to attack the question more directly and empirically. Are there methods that will ensure attention and focus on the critical features, whatever they may be?

Although little work in this area has been concerned with reproduction vs. recognition training, a recent theoretical discussion by Maccoby and Bee (1965) is relevant. These authors suggested that the subject must take account of more attributes of a form in order to reproduce it than to discriminate it from other forms. For example, children who can reproduce certain distinctive features of a form may fail to draw them in proper relation to each other, although they can recognize the form in an array perfectly. The authors propose that the lag in development between discrimination and reproduction may be accounted for by a difference in the number of attributes of the model necessary for the two tasks.

These comments suggest that reproduction training could produce better recognition of forms than explicit discrimination training by forcing the subject to attend to more criteial attributes. That is, if the subject must abstract more distinctive features in reproduction training in order to solve the training task, he then will have available more cues for the new discriminations presented in testing.

Furthermore, it seems reasonable to expect that whether or not reproduction training would force closer attention to the form would depend on the similarity of the standard and the other stimuli from which it must be distinguished. For example, if the subject must discriminate a circle from a triangle, he needs only to notice that one of the forms has an angle. If, however, his task is to distinguish an upright triangle from an inverted one, he is required to notice more. Hence the use of training items involving stimuli that are only minimally different will force the subject to attend closely and thereby focus on many attributes of the stimulus. In the process he will be likely to abstract certain features which are critical for its differentiation.

This experiment compares three conditions: (1) recognition training where the comparison stimuli are quite different from the standard; (2) recognition training where the comparison stimuli are transformations of the standards; and (3) reproduction training. It is predicted that (2) will be more effective than (1), because (2) forces more attention during training to the critical features of the forms. If reproduction training does in fact produce discrimination of more attributes than recognition training, then (3) will be most effective. If not, then reproduction training will be less effective than (2) and perhaps even than (1).

Method

Subjects: Subjects were kindergarten pupils enrolled in a Philadelphia elementary school. The basic experiment (Experiment 1), utilizing 32 of these children, was conducted during the first month of the kindergarten year. Another 32 subjects were run (Experiment 2) after they had completed 3/4 of their first school year. All subjects were between the ages four and one-half and five and one-half years old at the time of testing. An additional 11 children were dropped from the sample because they were uncooperative or because they were absent on the second day of testing. Subjects were assigned randomly to one of the four training conditions, with the restriction that the sexes be represented equally in each group.

Materials: The stimuli consisted of six nonsymmetrical letter-like forms, modified from those designed by Gibson et al. (1962) to follow the constraints of printed upper case capitals. Four transformations of these standard stimuli were also used: right-left reversal, up-down reversal (mirror image), 180° rotation, and 90° rotation. The standards and their transformations are shown in Figure 1. Twelve additional stimuli were also used, different from the six standards but constructed according to the same set of rules. These forms (individually and in combination) were printed in red on white 5" x 8" cards.

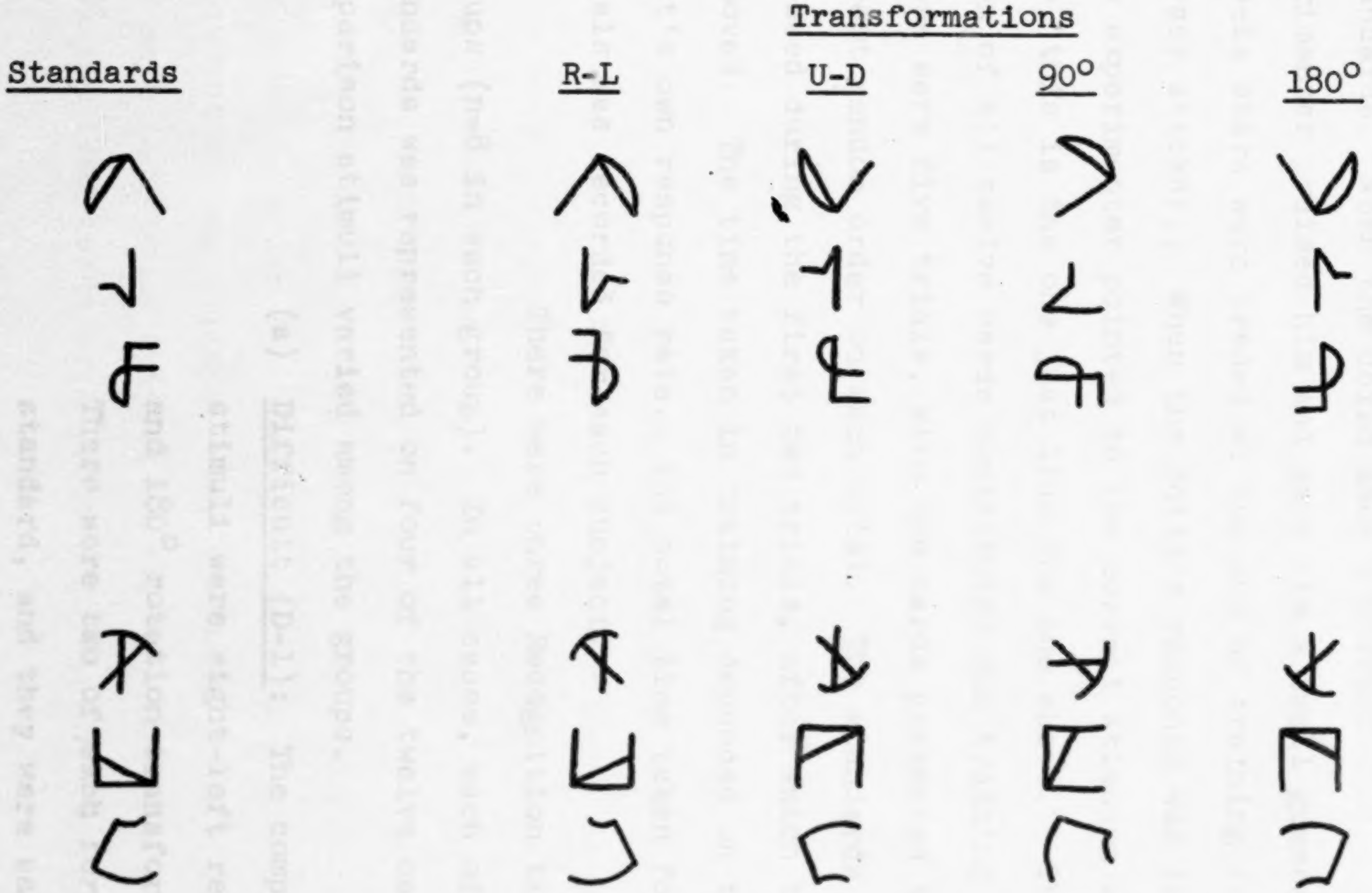
Training: The two basic training conditions were (1) Recognition, and (2) Reproduction training.

1. Recognition training:

The training task consisted of a delayed matching-to-sample task, in which the subject was asked to identify the standard after it was removed from view. This visual memory task was used because it approximates the perceptual learning tasks involved in actual reading more closely than does simple discrimination training (Wohlwill and Wiener, 1964).

Three of the six standards described above were placed in an array in front of the subject. (Half the subjects were shown three of the standards, chosen randomly from the six, and the other subjects were shown the other three.) Twelve cards, on each of which were two forms (one

Figure 1: The Standard Forms and their Transformations


Transformations
Standards
R-L
U-D
90°
180°

of the standards and another form) were presented in sequence, and the subject's task was to choose the form on each card which was exactly the same as one of the three standards. When the child made a correct choice, the experimenter praised him and gave him a small gummed star. (These stars were traded at the end of training for a larger sticker.) When the child's response was incorrect, the experimenter pointed to the correct stimulus and said, "No, this is the one just like the one above." The presentation of all twelve cards constituted one training trial. There were five trials, with the cards presented in a different random order on each trial. The standards remained exposed during the first two trials, after which they were removed. The time taken in training depended on the subject's own response rate. The total time taken for the five trials was recorded for each subject.

There were three Recognition training groups (n=8 in each group). In all cases, each of the three standards was represented on four of the twelve cards. The comparison stimuli varied among the groups.

(a) Difficult (D-1): The comparison stimuli were right-left reversal and 180° rotation transformations. There were two of each for each standard, and they were balanced as to whether they were placed to the right or the left of the standard on the card.

(b) Difficult (D-2): The comparison stimuli were up-down reversal and 90° rotation transformations, arranged as above.

(c) Simple (S): The comparison stimuli were four forms other than those used as the standards. Half the subjects in this group were shown one set of forms, and for the other half, a completely different set of comparison stimuli was used.

2. Reproduction training (R):

As in each of the Recognition groups, there were 8 subjects in this group. The three standards were presented in an array as in recognition training. However, no other forms were presented. The subject was asked to trace (twice) and copy (three or four times) each standard. The order in which the three standards were presented for reproduction was balanced over subjects. The training time for each individual subject in this condition was matched with the time taken by a subject in one of the recognition training groups, each of the three recognition training groups being represented equally. Approximately the same number of stars was given as in the other groups.

The order in which the three standards were placed in front of the subject at the beginning of training was balanced within and across groups. Half the Ss were given one set of three standards, and the other half were given the other three standards.

Experiment 2: The design of this experiment was similar to that described above, except that three, and not five, training trials were given. On the first trial, the standards were exposed, and on the other two they were hidden. This was the only modification in the design.

Testing: A series of three tests was administered immediately upon the completion of training, and the same series of tests was repeated 24 hours afterwards.

Test 1 (& Retest 1) consisted of a series of 24 cards (5" x 8") on each of which was drawn two forms, one standard and one other. Each of the three standards was presented four times with dissimilar forms (those used in S training) and once each with the four transformations. The cards were divided into two sets of twelve; and half the Ss received one set first and the other Ss, the other set first. The cards within each set were presented in a different random order to each subject. The subject was required to point to the standard on each card. No feedback was given on this or any of the tests.

Test 2 (& Retest 2) consisted of a series of 15 cards. On each card (8-1/2" x 11") were printed two sets of three forms each. One set contained one of the three standards. The other set contained a transformation of that standard. The other four forms, two in each set, had not previously been seen by the subject. Each standard was presented five times, once paired with each of the four transformations and once paired with one of the other forms used in Group S training. The standard and its comparison stimulus were always placed in the middle of their respective sets. The cards were presented in a different random order for each subject. The subject was required to point to the set which contained the standard.

Test 3 (& Retest 3) consisted of twelve cards (8-1/2" x 11"). On each card were placed two pairs of forms. One pair consisted of two of the three standards; the other pair consisted of one standard and a transformation of one of the other two standards. The cards were presented in a different random order for each subject. The subject was required to point to the pair which contained the two standards.

Results

A comparison was made of the number of errors made in training by each of the three recognition training groups. For the reproduction group (Group R), of course, it was impossible to make any comparable assessment of performance. In the first experiment, the mean number of errors in Group D-1 was 17.00; in Group D-2, 7.12; and in Group S, 0.50. The difference among these means was significant at the .01 level (F=11.95, df=2/18). Specific comparisons showed that Group S differed significantly from the other two (F=23.68, df=1/18, $p<.01$) and that Groups D-1 and D-2 also differed from each other (F=8.45, df=1/18, $p<.01$). The second experiment showed the same pattern; the mean errors for the three groups were as follows: D-1: 6.62, D-2: 2.00, S: 0.50 (F=5.09, df=2/18; $p<.05$). Again, Group S differed from the other two groups (F=7.26, df=1/18, $p<.01$), and Groups D-1 and D-2 differed from each other (F=5.34, df=1/18, $p<.05$).

The relative difficulty of the three tests was assessed by comparing the proportion of errors made on each of the three tests and their retests. Table 1 presents the mean proportion of errors for the two replications. The proportion of errors increased as a function of how much the test differed from the training task (F=6.07, df=2/380; $p<.01$). The performance in the second experiment was significantly superior to that of the first (F=144.13, df=1/190,

Table 1

Proportion of Errors on Each Test

	Test & Retest 1	Test & Retest 2	Test & Retest 3
Experiment 1	.29	.52	.68
Experiment 2	.16	.29	.39

$p < .01$), even though these subjects had received fewer training trials. There was no interaction between the two variables.

Performance on each of the three tests was analyzed as a function of the training conditions, the particular set of three standard stimuli to be discriminated, and the time of testing.

Experiment 1: On all three tests, differences among training methods were significant (Test 1: $F=4.27$, $p < .05$; Test 2: $F=9.57$, $p < .01$; Test 3: $F=3.27$, $p < .05$; all $df=3/24$). The two sets of standards did not differ (the respective Fs were 2.88, 2.11, and < 1, $df=1/24$). On Test 1, there were fewer errors on the retest than on the original test ($F=5.17$, $df=1/24$, $p < .05$), and there was a tendency for the retest to show fewer errors on the other two tests, although those differences did not reach significance. None of the interactions reached significance.

Experiment 2: There were no differences among training methods (Test 1: $F < 1$; Test 2: $F=1.53$; Test 3: $F=1.54$, $df=3/24$), nor were there differences in the standards used (the Fs were 4.19, 3.29, and < 1, $df=1/24$). As in the first experiment, there were fewer errors on the retest than on the original test in Test 1 ($F=5.07$, $df=1/24$, $p < .05$), and on the other two tests, the retest performance remained superior, though not significantly so (F=1.18 and <1, df=1/24). There were no significant interactions.

The next analysis concerned differences among training methods as a function of the type of transformation with which the standard was compared. Each of the three test and retest combinations was analyzed separately. Because of the significant differences between the two experiments, they were subjected to separate analyses. There were practically no errors on those comparison stimuli which were totally dissimilar from the standards and not simply transformations of them. Because these errors were so negligible, those items were excluded from the analyses. Tables 2 and 3 present the mean number of errors as a function of type of training and type of transformation. Tables 4 and 5 present summaries of the analyses of variance.

Experiment 1: Type of training was a significant variable, of course. Specific comparisons on Test 1 indicated that the two difficult training procedures were not different ($F<1$; df=1/28), but were significantly superior (i.e., showed fewer errors) to the simple discrimination and the reproduction training (F=12.90, df=1/28, $p<.01$). The simple recognition training and the reproduction training did not differ (F=1.23; df=1/28). Exactly the same pattern

Table 2

Number of Errors as a function of Type of Training and Type of Transformation (Experiment 1)

	D-1			D-2			Total	
	Test & Retest 1	Test & Retest 2	Test & Retest 3	Test & Retest 1	Test & Retest 2	Test & Retest 3		
R-L	20	20	21	14	17	23	115	R-L
U-D	9	15	16	10	9	15	74	U-D
90°	5	10	7	5	4	5	36	90°
180°	11	8	8	6	11	13	57	180°
		Total D-1=150			Total D-2=132			

	S			R			Total	
	Test & Retest 1	Test & Retest 2	Test & Retest 3	Test & Retest 1	Test & Retest 2	Test & Retest 3		
R-L	23	25	21	24	19	20	112	R-L
U-D	19	20	26	14	15	29	123	U-D
90°	21	19	18	16	16	7	97	90°
180°	16	21	21	11	20	20	109	180°
		Total S=230			Total R=211			

Note: Each figure represents the total number of errors for 8 subjects

Table 3

Number of Errors as a function of Type of Training and Type of Transformation (Experiment 2)

	D-1			D-2			Total	
	Test & Retest 1	Test & Retest 2	Test & Retest 3	Test & Retest 1	Test & Retest 2	Test & Retest 3		
R-L	17	13	22	19	15	19	105	R-L
U-D	3	6	8	2	4	7	30	U-D
90°	3	3	3	2	5	1	17	90°
180°	4	5	3	4	5	8	29	180°
		Total D-1=90			Total D-2=91			

	S			R			Total	
	Test & Retest 1	Test & Retest 2	Test & Retest 3	Test & Retest 1	Test & Retest 2	Test & Retest 3		
R-L	17	14	13	12	16	26	98	R-L
U-D	7	5	7	14	11	14	58	U-D
90°	6	7	3	10	9	4	39	90°
180°	3	5	11	4	12	13	48	180°
		Total S=98			Total R=145			

Note: Each figure represents the total number of errors for 8 subjects

Table 4

Analysis of Variance of Number of Errors on the Three Tests (Experiment 1)

Source	df	Test & Retest 1		Test & Retest 2		Test & Retest 3	
		ms	F	ms	F	ms	F
Type of Training (A)	3	12.21	4.92**	11.61	8.35**	8.20	3.78*
Error (b)	28	2.48	--	1.39	--	2.17	--
Transformation Type (B)	3	9.02	7.05**	5.65	3.12*	16.76	9.21**
A x B	9	1.04	<1	1.05	<1	2.41	1.32
Error (w)	84	1.28	--	1.81	--	1.82	--

*$p<.05$

**$p<.01$

Table 5

Analysis of Variance of Number of Errors on the Three Tests (Experiment 2)

Source	df	Test & Retest 1		Test & Retest 2		Test & Retest 3	
		ms	F	ms	F	ms	F
Type of Training (A)	3	1.19	<1	2.90	1.32	3.80	1.29
Error (b)	28	2.68	--	2.20	--	2.95	--
Transformation Type (B)	3	15.99	17.57**	8.22	7.98**	25.84	30.44**
A x B	9	1.76	1.93	0.29	<1	1.31	1.54
Error (w)	84	.91	--	1.03	--	0.85	--

$^{*}p < .05$

$^{**}p < .01$

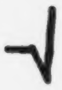
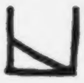
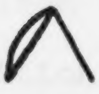
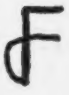
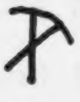
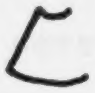
was seen on the other tests: *Test 2*: D-1 and D-2 *vs*. S and R: $F=20.91$, $p<.01$; D-1 *vs*. D-2: $F=1.62$; S *vs*. R: $F=2.26$. *Test 3*: D-1 and D-2 *vs*. S and R: $F=10.50$, $p<.01$; D-1 *vs*. D-2: $F<1$; S *vs*. R: $F<1$ (in each case $df=1/28$). On all three tests there was a tendency for Group S to show more errors in training than Group R, but this trend never reached significance. The other main effect, type of transformation, also reached significance on Test 1 and Test 3, although not on Test 2. (An analysis of the specific differences among the transformations is presented below.) There was no interaction between type of training and type of transformation.

Experiment 2: Type of transformation was a significant variable, but type of training did not reach conventional levels of significance. There were no interactions.

In order to assess the differences among the various transformations, a further analysis was done on the proportion of errors made on each of the four transformations, computed for each of the six standards individually. These data are presented in Table 6. While the order of the difficulty of the transformations depended on the particular stimulus, there was indeed a reliable difference among the

Table 6

Proportion of Errors on each Form and its Transformations

Standards	Total Errors (All Tests)	R-L	U-D	Transformations 90°	180°
	229	.24	.14	.26	.36
	223	.51	.19	.04	.26
	203	.34	.42	.19	.05
	200	.44	.27	.13	.16
	152	.51	.26	.10	.13
	123	.30	.15	.29	.26

transformation types ($F=5.09$, $df=3/20$, $p<.01$). Specific comparisons showed that the right-left reversal (R-L) was more difficult than the up-down reversal (U-D), the 90° rotation and the 180° rotation ($F=14.18$; $df=1/20$; $p<.01$). These other transformations did not differ among themselves.

Discussion

As predicted, recognition training in which the comparison stimuli were transformations was superior to recognition training where the comparison stimuli were totally different forms. This suggests that the comparisons involving minimally different stimuli did force the subject to attend to and abstract more attributes of the standard, which were then available for new test comparisons.

Reproduction training was not as effective as recognition-with-transformations, but was as effective as the simple recognition training. In fact, although the differences between reproduction training and simple recognition training were not significant, reproduction training scores were somewhat higher in both experiments. These results suggest that in reproduction training the subject was focussing on the standards to the same extent at least as in simple recognition training, though less than in the recognition-with-transformations training. Thus, in terms of the Gibson hypothesis, the number of attributes that will be abstracted by copying and tracing as compared to discrimination training does indeed depend on the similarity of the forms used in the discrimination training.

At first glance, it would appear that constructed-response training is not superior to multiple-choice training in this type of situation, as it is in continuous-discourse programs. However, the discrepancy between the earlier studies (Fry, 1960; Williams, 1963, 1965) and the present findings can be reconciled. In the present study, recognition training is actually a form of constructed-response training, for the very term "constructed-response training" implies a construction by the subject of the same type that will be expected on the test. In this experiment the behavior required on the test was that the subject choose one stimulus from an array, and therefore the training task was essentially the same as the test task. According to this interpretation, then, this experiment also shows the relative superiority of the constructed-response mode.

The greater difficulty of the right-left reversal corroborates the findings of Wohlwill and Wiener (1964), Rosenblith (1965) and, to some extent, Gibson et al. (1962). The difficulty of specific transformations varies as a function of the particular stimulus, however: transformations other than the right-left reversal were distinctly more difficult for two of the six standards. Gibson et al. (1962) have stated that the transformation types are more important as "predictors of identifiability" than are the characteristics of the standard itself. However, this conclusion does not seem warranted on the basis of the present study. It might prove instructive to analyze a set of stimuli in order to specify the variables that determine the order of difficulty of the transformations.

Overall performance in Experiment 2 was significantly better than that in Experiment 1, and in Experiment 2 there were no reliable differences among the training conditions. These differences are most likely due to the fact that the subjects in Experiment 2 had had several months of kindergarten experience, and for them, the training and/or testing conditions were too simple to show differences.

In both experiments there was a tendency for scores to be better on the retest than on the original test. This was probably because of additional learning which resulted from working through the original test itself. The difference between original testing and retesting reached significance in Test 1, the test most similar to the training task.

Even after a rather small amount of training, there were reliable differences among the training groups, and the data indicate that the effectiveness of readiness training does indeed depend on the particular techniques used. While special attention is given to practice on rotations and reversals in many readiness materials, usually too much time is devoted to copying and tracing or to discrimination exercises that are relatively ineffective, according to the present data. Moreover, such systematic training is sometimes given only in remedial work, that is, after a child has already developed some difficulty. The present experiment suggests (1) that more time be devoted to discrimination and recognition training that involves comparison of letters with their transformations, and (2) that this type of training be given early. It is clearly effective at the very start of kindergarten, and obviously, if appropriate training techniques and practice materials were used in the beginning stages of instruction, there should be less necessity later for remedial techniques.

Summary and Conclusions

The most effective method of training the recognition skills tested here was recognition training in which transformations of the forms were the comparison stimuli. In that this training task was highly similar to the final test task, it can be interpreted as essentially a constructed-response training mode. A group in which the comparison stimuli were completely different from the standards showed a lower error rate during training but much inferior test performance. It was hypothesized that the differences among types of training were due to differences in the subject's attention to the stimuli and differences in the number of criterial attributes focussed on and abstracted.

Performance level after reproduction training, in which the subject traced and copied the standard stimuli, fell between the transformation and non-transformation types of recognition training. The answer to the question of whether reproduction training or recognition training is more effective clearly depends on what kinds of comparisons are required during the recognition training.

The number of errors on the tests was a function of the type of transformation; right-left reversals were most difficult. However, the characteristics of the particular standard also determined to some extent the relative difficulty of its various transformations.

Since the methods of training showed such clear differences after a rather small amount of training, the following two suggestions seem warranted. First, the data indicate that systematic training of this kind can indeed be effective and efficient for children who are just beginning their schooling. It is recommended that such training be given as a matter of course, in order to help guard against the possible occurrence of certain relatively common perceptual confusions. Secondly, this experiment assesses a few of the basic techniques used in most of the reading readiness materials available today. It would appear that there are wide variations in the usefulness of typically-used materials, and the present data suggest that a relatively large amount of time might well be devoted to practice on discriminating letters (and perhaps other symbols) from their own transformations.

REFERENCES

Detambel, M. H., and Stolurow, L. M. Stimulus sequence and concept learning. Journal of Experimental Psychology, 1956, 51, 34-40.

Doty, B. A. & Doty, L. A. Programed instructional effectiveness in relation to certain student characteristics. Journal of Educational Psychology, 1964, 55, 334-338.

Evans, J. L., Glaser, R., & Homme, L. E. A preliminary investigation of variation in the properties of verbal learning sequences of the "Teaching Machine" type. In A. A. Lumsdaine & R. Glaser (Eds.), Teaching Machines and Programmed Learning. Washington, D. C.: National Education Association, 1960. Pp. 446-451.

Fernald, G. M. Remedial techniques in basic school subjects. New York: McGraw-Hill Book Co., 1943.

Ferster, C. B., & Sapon, S. M. An application of recent developments in psychology to the teaching of German. Harvard Educational Review, 1958, 28, 58-69.

Fitzgerald, H. Teaching machines: A demurrer. School Review, 1962, 70, 247-256.

Fry, E. B. A study of teaching machine response modes. In A. A. Lumsdaine & R. Glaser (Eds.), Teaching Machines and Programmed Learning. Washington, D.C.: National Education Association, 1960. Pp. 469-474.

Gibson, E. J. Learning to read. Science, 1965, 148, 1066-1072.

Gibson, E. J., Gibson, J. J., Pick, A. D., and Osser, H. A developmental study of the discrimination of letter-like forms. Journal of Comparative and Physiological Psychology, 1962, 55, 897-906.

Gibson, E. J., Osser, H., Schiff, W., & Smith, J. An analysis of critical features of letters tested by a confusion matrix. A Basic Research Program on Reading. U.S. Department of Health, Education and Welfare, Office of Education, Cooperative Research Project No. 639. Ithaca, N. Y.: Cornell University, 1963. Unpaged.

Gleason, G. Programmed learning in sixth grade arithmetic. Paper presented at American Educational Research Association, Chicago, February 1964.

Goldbeck, R. A. & Campbell, V. N. The effect of response mode and response difficulty on programed learning. _Journal of Educational Psychology_, 1962, _53_, 110-118.

Hershberger, W. Self-evaluational responding and typographical cueing. _Journal of Educational Psychology_, 1964, _55_, 288-296.

Holland, J. G. Design and use of a teaching machine program. Paper read at American Psychological Association, Chicago, September 1960. (Mimeo)

Holland, J. G., & Porter, D. The influence of repetition of incorrectly answered items in a teaching machine program. _American Psychologist_, 1960, _15_, 423.

Keislar, E. R. & McNeil, J. D. A comparison of two response modes in an autoinstructional program with children in the primary grades. _Journal of Educational Psychology_, 1962, _53_, 127-131.

Krumboltz, J. D. & Weisman, R. G. The effect of overt versus covert responding to programed instruction on immediate and delayed retention. _Journal of Educational Psychology_, 1962, _53_, 89-92.

Lambert, P., Miller, D. M., & Wiley, D. E. Experimental folklore and experimentation: The study of programmed learning in the Wauwatosa public schools, Journal of Educational Research, 1962, 55, 485-494.

Maccoby, E. E. and Bee, H. L. Some speculations concern-the lag between perceiving and performing. Child Development, 1965, 36, 367-377.

Moore, O. K. Autotelic responsive environments and exceptional children. Hamden, Conn.: Responsive Environments Foundation, 1963.

Pick, A. D. Improvement of visual and tactual form discrimination. Journal of Experimental Psychology, 1965, 69, 331-339.

Popp, H. M. Visual discrimination of alphabet letters. Reading Teacher, 1964, 17, 221-226.

Porter, D. Some effects of year long teaching machine instruction. In E. H. Galanter (Ed.), Automatic teaching: The state of the art. New York: Wiley, 1959. Pp. 85-90.

Rosenblith, J. F. Perceptual development in normal children and others. Unpublished monograph, 1965.

Smith, N. B. Matching ability as a factor in first grade reading. Journal of Educational Research, 1928, *18*, 560-571.

Spalding, R. B., & Spalding, W. T. *The writing road to reading*. New York: Whiteside, 1957.

Stolurow, L. M. Teaching by machine. *Cooperative Research Monograph No. 6*. Washington: United States Government Printing Office, 1961.

Taber, J. I., Glaser, R. & Schaefer, H. H. *Learning and programmed instruction*. Reading, Mass.: Addison-Wesley Co., 1965.

Vernon, M. D. *Backwardness in reading*. Cambridge, England: Cambridge Univer. Press, 1957.

Williams, J. P. Comparison of several response modes in a review program. *Journal of Educational Psychology*, 1963, *54*, 253-260.

Williams, J. P. Effectiveness of constructed-response and multiple-choice programing modes as a function of test mode. *Journal of Educational Psychology*, 1965, *3*, 111-117.

Williams, J. P. & Levy, E. I. Retention of introductory and review programs as a function of response mode. *American Educational Research Journal*, 1964, *1*, 211-218.

Wohlwill, J. F. and Wiener, M. Discrimination of form orientation in young children. *Child Development*, 1964, *35*, 1113-1125.